Homework #3

See Canvas for HW #3 assignment due date.

This week, there won't be any "theory" homework. Please use that time to work on your project proposal.

Problem B.1: Model Selection Criterion

In this lesson, we will perform both the full and partial F-tests in R.

We will use the Amazon book data.

https://raw.githubusercontent.com/bzaharatos/-Statistical-Modeling-for-Data-Science-Applications/master/Modern Regression Analysis /Datasets/amazon.txt

The data consists of data on $n = 325$ books and includes measurements of:

- `aprice` : The price listed on Amazon (dollars)
- `lprice` : The book's list price (dollars)
- `weight` : The book's weight (ounces)
- `pages` : The number of pages in the book
- `height` : The book's height (inches)
- `width` : The book's width (inches)
- `thick` : The thickness of the book (inches)
- `cover` : Whether the book is a hard cover of paperback.
- And other variables...

I will include some data cleaning to get you started, although you don't have to use this exact code. We do want to remove NA and average out what we can beforehand.

For all tests in this lesson, let $\alpha = 0.05$.

```
## Here is the data cleaning I mentioned. Again, feel free to explore this via your own method.
paste0("https://raw.githubusercontent.com/bzaharatos/",
                    "-Statistical-Modeling-for-Data-Science-Applications/",
                    "master/Modern%20Regression%20Analysis%20/Datasets/amazon.txt")

df = data.frame(aprice = amazon$Amazon.Price, lprice = as.numeric(amazon$List.Price),
                pages = amazon$NumPages, width = amazon$Width, weight = amazon$Weight..oz,
                height = amazon$Height, thick = amazon$Thick, cover = amazon$Hard..Paper)

df$lprice[which(is.na(df$lprice))] = mean(df$lprice, na.rm = TRUE)
df$weight[which(is.na(df$weight))] = mean(df$weight, na.rm = TRUE)
df$pages[which(is.na(df$pages))] = mean(df$pages, na.rm = TRUE)
df$height[which(is.na(df$height))] = mean(df$height, na.rm = TRUE)
df$width[which(is.na(df$width))] = mean(df$width, na.rm = TRUE)
df$thick[which(is.na(df$thick))] = mean(df$thick, na.rm = TRUE)
```

B.1. (a) The Model

We want to determine which predictors impact the Amazon list price. Begin by fitting the full model.

Fit a model named `lmod.full` to the data with `aprice` as the response and all other rows as predictors. Then calculate the AIC, BIC and adjusted R^2 for this model. Store these values in `AIC.full`, `BIC.full` and `adj.R2.full` respectively.

B.1. (b) A Partial Model

Fit a partial model to the data, with `aprice` as the response and `lprice` and `pages` as predictors. Calculate the AIC, BIC and adjusted R^2 for this partial model. Store their values in `AIC.part`, `BIC.part` and `adj.R2.part` respectively.

B.1. (c) Model Selection

Which model is better, `lmod.full` or `lmod.part` according to AIC, BIC, and R_a^2? Note that the answer may or may not be different across the different criteria. Save your selections as `selected.model.AIC`, `selected.model.BIC`, and `selected.model.adj.R2`.

B.1. (d) Model Validation

Recall that a simpler model may perform statistically worse than a larger model. Test whether there is a statistically significant difference between `lmod.part` and `lmod.full`. Based on the result of this test, what model should you use?

Problem B.2

`divorce` is a data frame with 77 observations on the following 7 variables.

1. `year` : the year from 1920-1996
2. `divorce` : divorce per 1000 women aged 15 or more
3. `unemployed` unemployment rate
4. `femlab` : percent female participation in labor force aged 16+
5. `marriage` : marriages per 1000 unmarried women aged 16+
6. `birth` : births per 1000 women aged 15-44
7. `military` : military personnel per 1000 population

Here's the data: (I'll also include all data links in Canvas)

```
paste0("https://raw.githubusercontent.com/bzaharatos/",
                     "-Statistical-Modeling-for-Data-Science-Applications/",
                     "master/Modern%20Regression%20Analysis%20/Datasets/divusa.txt")
```

B.2 (a) Using the `divorce` data, with `divorce` as the response and all other variables as predictors, select the "best" regression model, where "best" is defined using AIC. Save your final model as `lm_divorce`.

B.2 (b) Using your model from part (a), compute the variance inflation factors VIFs for each $\widehat{\beta}_j, j = 1, \dots, p$. Store them in the variable `v`. Also, compute the condition number for the design matrix, stored in `k`. Is there evidence that collinearity causes some predictors not to be significant? Explain your answer.

B.2 (c) Remove the predictor with the highest VIF. Does that reduce the multicollinearity?